Design Choices for Automated Disease Surveillance in the Social Web

Mark Abraham Magumba[1]*, Peter Nabende[1] and Ernest Mwebaze[2]

1. Department of Information Systems, Makerere University Uganda, College of Computing and Information Sciences

2. Department of Computer Science, Makerere University Uganda, College of Computing and Information Sciences

Abstract:

The social web has emerged as a dominant information architecture accelerating technology innovation on an unprecedented scale. The utility of these developments to public health use cases like disease surveillance, information dissemination, outbreak prediction and so forth has been widely investigated and variously demonstrated in work spanning several published experimental studies and deployed systems. In this paper we provide an overview of automated disease surveillance efforts based on the social web characterized by their different high level design choices regarding functional aspects like user participation and language parsing approaches. We briefly discuss the technical rationale and practical implications of these different choices in addition to the key limitations associated with these systems within the context of operable disease surveillance. We hope this can offer some technical guidance to multi-disciplinary teams on how best to implement, interpret and evaluate disease surveillance programs based on the social web.

Keywords: Crowd sourced Disease surveillance, Data Mining, Knowledge Engineering, Participatory Epidemiology, The social web

*Correspondence: magumbamark@hotil.com

DOI: 10.5210/ojphi.v10i2.9312

Introduction

Current global trends like increasing population densities and higher mobility of persons and goods mean epidemics have the potential to spread extremely rapidly [1,2] thereby creating a need for equally fast reporting for early detection and investigation of outbreaks. Whereas traditional approaches like sentinel surveillance are still relevant, they are incapable of adequately addressing the needs of early detection due to time lags introduced by formal processes and their limited geographical reach. Web-based systems have emerged as an addition to these efforts. Web-based approaches offer almost instant reporting which in the case of the social web is made possible by billions of users spread out across the globe and generating a continuous deluge of information in what is probably the largest data gathering operation in the world. Most of this data is in the form of natural language text and is devoid of structure necessitating information extraction routines before it can be converted to actionable epidemiological intelligence. It is also generated by non experts thereby raising legitimate concerns about its accuracy, reliability and verifiability. However, even long established approaches like EHR (Electronic Health Reporting) suffer from similar problems in addition to the fact that even in highly developed countries like the United States there are many areas where they are not well deployed [3-5].

Disease surveillance attempts to answer five questions about disease events of interest often abbreviated as the 5Ws. These are what (the event diagnosis),where (the event location), when (the time of occurrence), who (the victim/patient) and why/how (the causal agents involved). Disease surveillance requires a continuous and rigorous data acquisition operation which can be both complex and logistically demanding. Regarding the amount of directed effort expended in the process of obtaining information, surveillance can be characterized as either active or passive. With active surveillance those in charge of the process actively seek information as opposed to passive surveillance where medical workers wait for cases to essentially report themselves in the normal process of patients seeking medical care or some alternative scenario. Both approaches are feasible on the social web.

Disease surveillance can also be classified as indicator based or event based. Indicator based surveillance encompasses traditional formal systems with regular, predetermined reporting times whereas event based surveillance is real-time and ad hoc and incorporates both formal and informal sources and usually entails loose case definitions. The focus of Web-based surveillance has nearly exclusively been on event based surveillance and all the systems covered in this paper are event based systems. Another term that has become nearly synonymous with this web-based disease surveillance is syndromic surveillance which is defined as an investigational approach where health department staff, assisted by automated data acquisition and generation of statistical alerts, monitor disease indicators in real-time or near real-time to detect outbreaks of disease earlier than would otherwise be possible with traditional public health methods [1]. The goal of syndromic surveillance is early detection and it generally employs pre-diagnostic data or more relaxed definitions of incidence referred to as syndromes. In both cases disease events are considered unconfirmed and warrant further investigation. This is the stance adopted by most implementations based on the social web.

The social web has been formally defined as a set of social relations that link people through the World Wide Web [6]. In functional terms it encompasses several services such as microblogs (Twitter, Tumblr),social networking (Facebook, Google+, Linkedin, Whatsapp), video sharing (Youtube, Vimeo), image sharing (Instagram, Flickr) and blogging to mention but a few. However, the boundaries of the social web have been extended by the rise of social Application Programming Interfaces (APIs) which are programming libraries that allow third party applications to make function calls to web services, and it now potentially includes the entire World Wide Web. Through social APIs services that aren't already social are easily "socializable" by incorporating social features like content sharing and user comments hosted on third party platforms. An example is here is Google which has incorporated social services like Google Hot Trends, Google Social Search and Google+ to mention but a few. Similarly Chinese search giant Baidu implemented Baidu Space which allows registered users to create personalized homepages in a query-based searchable community and Baidu bookmarks a social bookmarking service supported by Baidu.com.

The vast reach of the social web whose users currently number in billions [7] and the rise of social platforms means social data is probably the most complete record of current human affairs. It does not just offer the possibility of capturing users views on a multitude of subjects but their physical and mental situation, geo-location, movements, connections and much more in a variety of data formats ranging from text to streaming video. This information is mostly unsolicited and is generated at an explosive pace with unprecedented variety. For instance Facebook.com, a single social networking service had 2.07 billion monthly active users of which 1.74 billion were daily active users generating information at a rate of 293,000 status updates, 510,000 comments and 136,000 photo updates a minute as of November 2017 [8]. This makes the social web a uniquely well suited architecture for the implementation of Web-based, syndromic surveillance systems. The key theme is user generated content and user interaction and by this criterion we shall exclude systems like Medisys [9] and GOARN (Global Outbreak and Alert response Network) [10] in which there is a strong distinction between the system's primary data entrants and the general public and the direction of information flow is strictly from the system to the general public with the general public being purely a consumer rather than a source of information.

Design Choices for Automated Disease Surveillance Systems and Studies on the Social Web

In this section we discuss some of the key high level design choices for disease surveillance systems and experimental studies. Figure 1 below depicts the typical analytical pipeline for Web-based disease surveillance. Dashed lines imply optional steps. Based on these steps we identify the following high level decisions that have to be made: The user participation model, language parsing methodology, multiplicity of data sources (whether single or multiple data sources are employed), number of diseases to be investigated, the ultimate objective (predictive modeling, real-time disease monitoring or explanatory modeling) and the choice of deployment platform. The simplest pipeline involves retrieving messages from a single source and aggregating and reporting this data usually as a map visualization. This applies to the case of explicit user participation where structured data is obtained from willing volunteers and disease reports are assumed to be true meaning there is no need for mathematical modeling or data preprocessing steps like translation, filtering and language parsing.

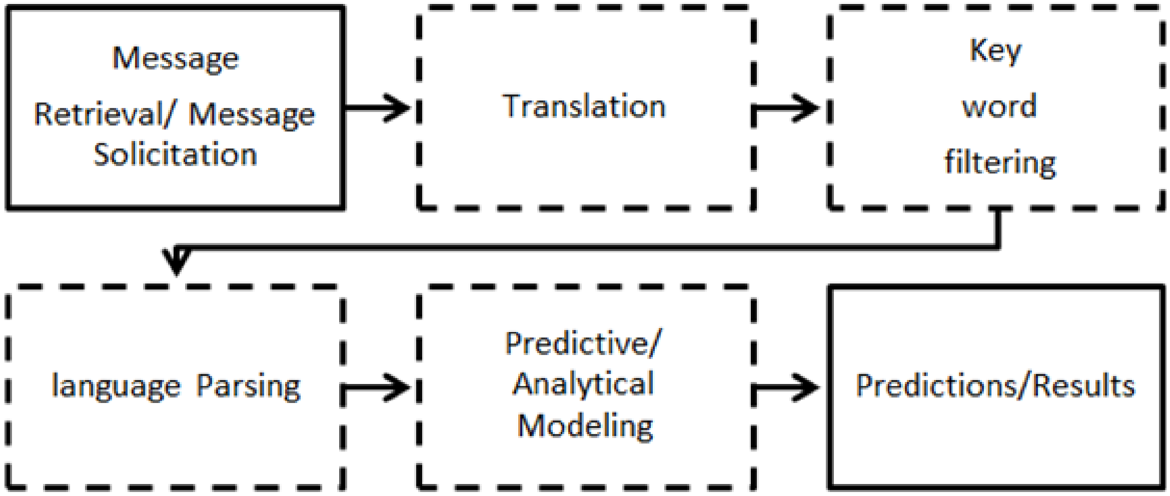
Message Retrieval/ Message Solicitation
Translation
Key word filtering
language Parsing
Predictive/ Analytical Modeling
Predictions/Results

Figure 1: Analytical Pipeline for Disease Surveillance applications on the social web

User Participation Model

Although the social web is participatory by nature, the extent to which users may be cognizant of the purposes for which their data is used differs. In most cases users are not explicitly recruited to volunteer information for the sole purpose of public health monitoring. For instance Google Flu trends and Google Dengue trends Google [11] which relied on user searches for flu and dengue fever related terms did not provide explicit notification to users that their searches were being monitored for this purpose. In this case it is assumed users have provided implicit permission for use of their data for additional purposes by agreeing to terms and conditions of social web services usually by the very act of using them. This approach raises some privacy concerns and some services like Twitter have attempted to offset them by allowing users to label their messages as private to make them unavailable for these additional purposes however the default visibility is set to public meaning user messages are available to third party access unless explicitly protected by the authors. This category also entails health news aggregators like GPHIN (Global Public Health Network) [12] and HealthMap.org which compile health related reports from a multitude of web based sources not necessarily dedicated to health reporting.

Another instance of implicit participation involves the use of automatically obtained data from embedded sensors like Propeller Health's sensor for asthma and COPD (Chronic Obstructive Pulmonary Disease). Propeller Health has used data from its network of asthma and COPD patients to produce so-called asthma forecasts as well as track local asthma and COPD conditions in real-time for American cities. The service known as "Propeller Air" has been packaged into an API allowing it to be integrated into other websites and web services [13-15]. It has even been integrated into Alexa, Gogle's AI assistant allowing users to obtain local asthma data via a natural language interface by simply saying, "Get asthma conditions from Propeller".

Implicit user permission appears to be the de facto stance of most systems and studies based on social media [16-22]. In this case users inadvertently report ongoing cases of interest in the course of using the system for some other purpose. Where unstructured textual data is involved this is technically challenging as it requires advanced text pre-processing and language parsing to extract useful information as messages are not necessarily generated for the purpose of communicating disease occurrence.

Alternatively, information may be explicitly requested for from the public. This is the approach adopted by implementations like flunearyou.org [23] for the United States of America which boasts over 60,000 users and influenzanet.eu [24] which currently covers eleven European countries. These systems rely on volunteers who knowingly and voluntarily report their flu symptoms. Influenzanet.eu started off as a Dutch service called degrotegriepmeting.nl which was able to recruit 20,000 volunteers in the Netherlands and Belgium in its first season and 30,000 volunteers in 2003 its first year of operation through a vigorous marketing campaign [25] and the so called Great Dutch influenza survey ("de Grote Griepmeting" in Dutch) has been carried out since then [26], the second service was gripenet.pt a Portuguese service that was directly inspired by the success of the Dutch system and attracted 5000 volunteers between 2005 and 2007 [27], a third service influweb.it was launched in Italy in 2008 [28] and finally a fourth system GrippeNet.fr was launched in France in 2012 and currently has 11811 active accounts [29]. These four teams decided to collaborate on the development of a unified European system directly leading to the birth of influenzanet.eu. In 2009, the Northern hemisphere H1N1pandemic accelerated plans to export the platform to a new country namely the United Kingdom in the form of a fifth system flusurvey.org.uk [30,31]. The influenzanet.eu project has been extended as influensakoll.se in Sweden [32], flusurvey.ie in Ireland [33], influmeter.dk in Denmark [34] and gripenet.es in Spain [35].

The appeal of explicit user involvement is that it requires less advanced language processing and is likely to generate better quality data than implicit user for instance ProMed-Mail [36] employs volunteer networks of professionals that report high quality disease incidence data for early detection of disease outbreaks with high accuracy. That said there is currently no conclusive research that consciously reported data from volunteers is better at predicting or monitoring diseases than automatically mined data inadvertently supplied by unwitting users in their public social posts. On the basis of the literature we have come across, implementations and experiments based on an implicit user participation model such as those conducted on Google Flu Trends and Twitter data [16,17,19,37] seem to slightly outperform those based explicit volunteer reporting [26,38,39] in terms of how well they correlate with official data. This is probably as a consequence of the fact that systems based on widely deployed social platforms like Twitter make up for their poor data quality by accessing a far larger number of potential “volunteers” comprising every one of the billions of social web users. Furthermore, even with self-reported data problems of verifiability persist. For instance it is not possible to remotely diagnose an infectious disease and services like influenzanet.eu have to rely on so-called “syndrome definitions” which are basically groups of symptoms that if they co-occur have a high probability of being indicative of a given illness. However, even when individuals faithfully report these, there is a significant overlap between different diseases in terms of symptoms and the same symptoms could be caused by potentially unrelated conditions.

Furthermore, systems based on self-reported information are useless when users do not provide this information for any reason and therefore require elaborate marketing campaigns and reminder mechanisms to ensure a continuous stream of data. Volunteer driven systems generally suffer from user attrition [40] and as a consequence require a concerted effort to maintain user participation and engagement in addition to continuous user recruitment. However, there is always a chance that these same users are inadvertently revealing this same information in the process of communicating with their social connections hence the utility of systems that are always listening in on user conversations for any relevant mentions of disease related words.

Finally, from an operational overhead point of view implicit user participation is far cheaper than explicit participation as social APIs eliminate the need to explicitly enroll volunteers allowing operational details like marketing and maintaining user engagement to remain external to the surveillance process. In addition by comparison to explicit participation it has very low infrastructure requirements as there is no technical need to store disaggregated user data. There are also implementations which employ a hybrid user participation for certain aspects of the process such as Crowdbreaks.org [21]which employs volunteers to annotate messages for its classification a mechanism they refer to as "user-driven data refinement".

Language Parsing Methodology

Where users are not explicitly aware that their conversations are being monitored for diseases surveillance messages will be highly unstructured and contain additional artifacts like emoticons, URLs (Universal Resource Locators), slang and multimedia in addition to errors like misspellings. Hence some language parsing is required and from this perspective web based disease surveillance systems can be broadly categorized as rule-based systems or machine learning systems. Rule-based approaches extract useful data by using manually generated rules. For instance Doan et al [41] use rules like the presence of references to negation and humour in conjunction with shallow parsing with the RASP grammatical parser [42] to filter Twitter messages about the flu. The simplest rule based systems employ keyword matching approaches which assume messages to be relevant if they contain certain keywords [11,19,43,44]. In projects like BioCaster [45], Bio Storm (Biological Spatio-Temporal Outbreak Reasoning Module) [46], MediSys [47], GPHIN and HealthMap.org [48] these keywords are have further been hierarchically structured into ontologies or taxonomies. Ontologies allow for more structured definitions that may simplify tasks like dictionary maintenance and support for multiple languages by modularizing concept definitions.

There are two approaches for implementing multi-lingual systems the first involves translating messages to a common language such as English and then applying some parsing rules that are usually based on an ontology, in this case you have a single ontology definition. This is the approach taken by the HealthMap project which relies on the Google Translate service for translation. The second involves creating different versions of the ontology for different languages which was the approach that was taken by the Biocaster project. The advantage of the former approach is that it is likely to support more languages as freely accessible public APIs are well maintained and constantly adding new languages for instance Google Translate currently supports 103 languages [49] whereas BioCaster supported 8 languages [50].The advantage of the latter approach is that it may provide better performance than machine translation which is still relatively inaccurate particularly for noisy data sources like social media and may not always produce reliable results.

However, it is difficult to create and maintain exhaustive domain definitions using static rules. Simple rules like text matching are susceptible to false positives as in many cases keywords may be mentioned in irrelevant contexts on the other hand longer more elaborate rules may be susceptible to false negatives as a result of the fact that it is impossible to write rules that account for the full variety of expressions people may use to communicate morbidity. This may result in inaccurate models as generally some positive correlative relationship is sought between the volume of keywords and disease activity on the ground. For instance equation 1 below depicts the formula employed to relate keyword volume with official ILI data by fitting the log-odds of an ILI physician visit with an ILI related search query for the Google Flu Trends project by Ginsberg et al [37]. P represents the percentage of ILI physician visits, β_0 is the intercept.

$$logit(P) = \beta_0 + \beta_1 \times logit(Q) + \varepsilon \; (1)$$

Given such a formulation, imprecise semantic filtering will likely negatively impact results. We are not aware of any studies that have attempted deep semantic filtering on search engine data but for microblog data such as Twitter data it has been shown to be the case that more robust semantic filtering improves results. This is achieved with a second approach known as machine learning. With a machine learning approach explicit rules do not need to be written as these are automatically learned from data by some machine learning algorithm. With this approach first a keyword-based approach is applied and then messages are further filtered using machine learning algorithms. Machine learning will generally impose higher infrastructure requirements due to the increased computational overheads of running machine learning algorithms.

Machine learning also imposes higher operational overheads as a result of the additional cycle of activities related to training and keeping language models updated. Figure 2 depicts a generic Natural language Processing (NLP) cycle for unstructured textual data typical of the social web. The NLP cycle starts with corpus generation which entails obtaining messages from which to create language models from social platforms. The key concern at this phase is ensuring that the messages obtained are actually representative of the data. Additionally obtaining a sufficiently dense sample may not be straightforward. Whereas there is an unprecedented amount of information on the social web, its distribution is tilted towards certain popular topics. Therefore for some topics it may not be possible to obtain sufficient information to create reasonable data driven language models. However, many social platforms that have implemented APIs simplify this step. For instance the Twitter API allows one to obtain randomly sampled messages for a given keyword eliminating the need to create web scrapping tools to obtain messages.

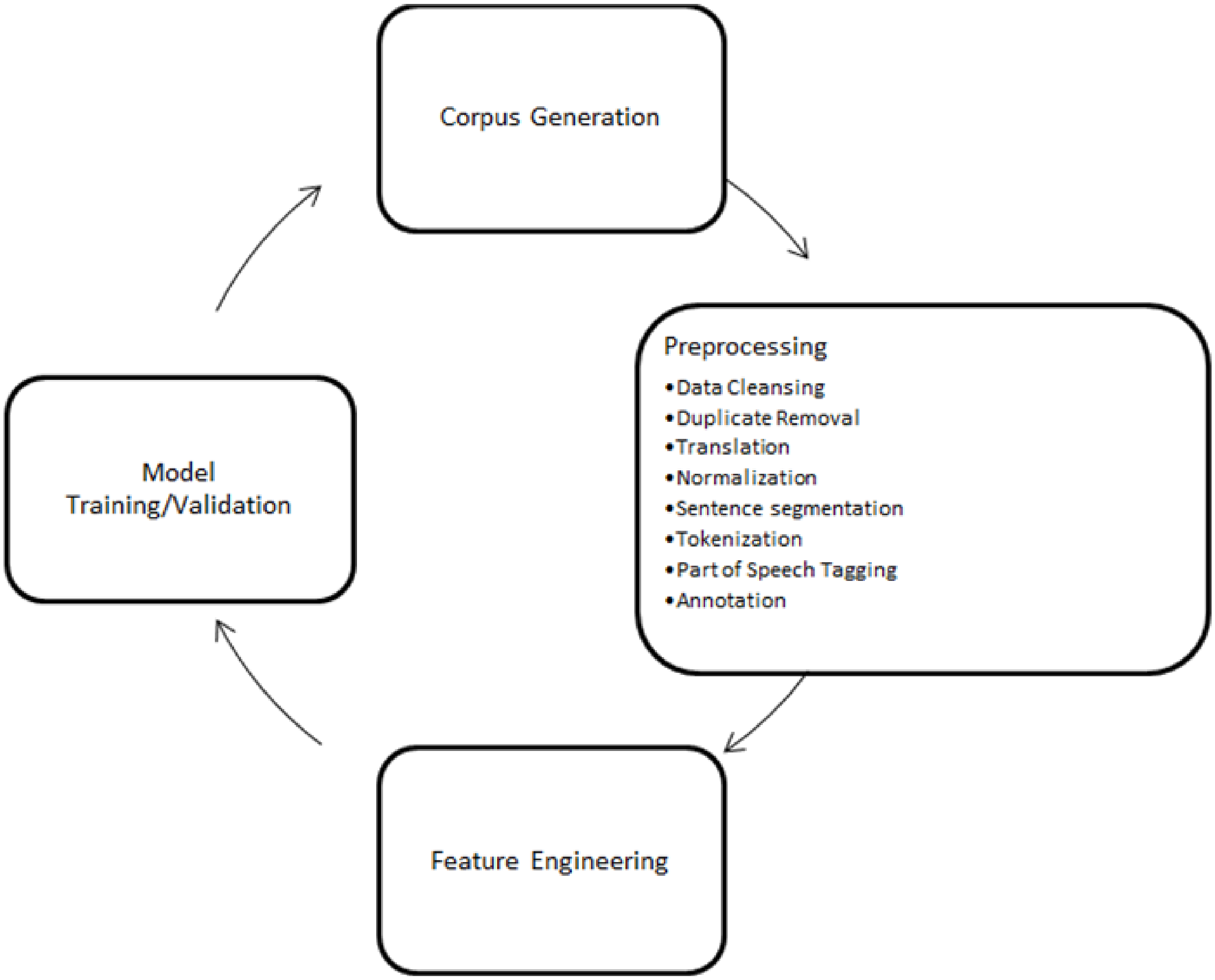

Figure 2: Activity cycle for textual Analytical pipeline for disease surveillance applications on the Social Web

The data also requires pre-processing which may involve several steps such as translation for multi-lingual pipelines. Other typical steps include duplicate removal, data cleansing steps like removal of punctuation and other social media related artifacts like hashtags, URLs and emoticons, sentence segmentation, tokenization, normalization, part of speech tagging, noun phrase chunking and finally annotation. The next step is feature engineering. In many situations the words themselves are taken as the features and simply vectorized. More advanced techniques are possible like distributional word embeddings such as neural word embeddings like word2vec [51] which represent words as low dimensional vectors and are able to conserve non trivial semantic relationships for instance where v(King) is the vector for “king” it has been found that v(King) – v(Man) + v(Woman) returns roughly the same value as v(Queen) . In addition distributional representations have been found to improve the results of NLP tasks such as classification and achieve state of the art results for other tasks like named entity recognition with methods such as recurrent neural networks. However, their application in this domain is limited to a few examples like the work by Magumba et al [52] who employed word2vec embeddings learned from conceptual representations of tweets to create message classification models for general purpose disease incidence detection.

The next step is training a model using these features followed by validation of the model. The entire sequence of activities required to encode a message as a vector of features has got to be repeated on the fly for every message for live systems. This includes all pre-processing and feature extraction steps in addition to decoding the message into some result using the learned model. Intermediate steps like part of speech tagging and noun phrase chunking have their own pipelines and require similar language resources like annotated training data for model training and validation.

Given the high throughput of these systems the computational requirements of this pipeline are a real technical concern that must be addressed at design time. For text processing algorithms computational complexity is typically expressed in terms of processing time (time complexity) and memory requirement (space complexity) as a function of some property of the data. Typical properties employed for complexity evaluations include the average length of a message and the number of messages. Usual workarounds to computational complexity typically rely on parellization strategies in which the computing load is distributed across multiple elements like processors which undoubtedly introduces additional costs. In addition algorithmic compromises are usually made for complex algorithms like CRFs and log linear models such as limiting the number of context words at a small cost to performance. Furthermore approximate inference approaches like Stochastic Gradient Descent (SGD) and Coordinate Descent can be employed to dramatically reduce the training time complexity of several methods such as logistic regression and SVMs (Support Vector Machines). Typical applications of machine learning in this domain include classification and named entity recognition. Classification entails distinguishing between relevant and irrelevant messages with classification algorithms like logistic regression classifiers [16], Support vector machines (SVMs) with polynomial kernels [18], log linear models [53], and Naive Bayes classification [54]. Named entity recognition (NER) which entails automatic detection of entities like persons, disease events and locations with methods such as SVMs [54] and linear chain Conditional Random Fields (CRFs) [55].

Unsupervised approaches on the other hand do not require any annotated data and instead rely on some objective function like the Euclidean distance between vectorial representations of messages. These techniques are generally clustering algorithms that partition a corpus of messages and other data into clusters of closely related signals. The clusters may be used as target categories to train classification models for classifying new messages or may be employed as syndroms for generalized syndromic surveillance as with Ailment Topic Aspect Model (ATAM) [56] employed in Healthtweets.org [20]. Other techniques include Latent Dirichlet Allocation (LDA) [57], Expectation Maximization (EM) clustering [58] and bisecting k-means algorithm [59] to mention but a few.

These techniques will also require some feature engineering to overcome noise like irrelevant words and achieve some level of generalization. Purely data driven approaches commonly employ a vector of the most frequently occurring words as the vocabulary and mask over rare words. The size of the vocabulary is usually set by trial and error. A thresholding approach where tokens whose frequencies are below a certain value are eliminated may also be used as an alternative to setting some arbitrary vocabulary length. Depending on the algorithm the resulting representation may be the same length as the original message as with distributional clustering approaches like LDA or a fixed length vector with the same length as the vocabulary as with centroid based clustering techniques like k-means. In the latter case a one hot encoding is typically employed which is a binary representation indicating whether or not each vocabulary item occurs in a given message. These datasets are usually characterized by large vocabularies and short messages therefore this is usually a very sparse vector with most items set to zero. This can then be subjected to additional transformations like Principal Component Analysis for dimensionality reduction [58].

The key difficulty with unsupervised methods is that they require one to set hyper parameters and choosing optimal values for hyper parameters could be challenging and requires one to make some potentially unjustified a priori assumptions about the data like the number of topics in addition to alpha and beta parameters that describe the distribution of words over topics and topics over documents respectively for LDA. Considering that word and topic distributions are not static on the social web this raises practical concerns on how to keep models up to date as good guesses of these hyper parameters have to be continuously supplied as the underlying word and topic distributions change. Furthermore, sometimes there are issues of interpretability of results for instance LDA will not always return topics that make sense and therefore require some human mediation.

Finally there are implementations that employ a hybrid human mediated automatic language parsing procedure such as GPHIN which automatically calculates so-called “relevancy scores” for each article. Articles with a high score are automatically published but those with low scores go through a second manual check by a human analyst [60]. Such a setup minimizes the possibility of false negatives.

Multiplicity of Data Sources

It is a requirement for these approaches to analyze as much signal data as possible in order for them to be as representative as possible. The natural approach to achieving this has been to incorporate as many data sources as possible. This is an the approach that has been employed by systems like Healthmap.org and the “Outbreaks Near Me” mobile application which incorporate data from GoogleNews, ProMed-Mail, World health organization reports, Geo Sentinel, OIE (the World organization for animal health), EuroSurvillence, Baidu News, Soso info and the Wildlife Data Integration Network. Many of these sources are themselves aggregators of multiple sources for instance Google News aggregates more than 50,000 news websites [61], ProMed-Mail itself relies on media reports, official reports, online summaries, local observers, and others.

Whereas there is some experimental evidence that suggests that employing multiple data sources may improve accuracy and robustness of predictive algorithms, maintaining multiple APIs and data extraction pipelines imposes a significant amount of operational overhead on the process. There also additional issues that will have to be resolved such as dealing with different data representations and formats and varying spatial and temporal granularity across different data sources.

However, the rise of social media and social APIs has made single source surveillance feasible. Social networking websites like Twitter.com and Facebook.com are in fact aggregators in away as they allow multiple types of users to sign on. There are three key types of users. The first are individuals who number in billions. The second are organizations like news services and government agencies communicating vital information like updates on developing situations in real time. The third category comprises automated agents known as bots. For this reason social media offers a simplified route to achieving representative data source as it bypasses the complex logistics of maintaining multiple APIs and data sources. Single source systems include examples like BioCaster which employed ProMed-Mail alerts, flu-prediction.com, Crowd-breaks.org and HealthTweets.org which are based on Twitter data in addition to most studies based on Twitter [16,17,19,55,62-64], Polgreen's work on Yahoo searches [39] and the now retired Google Flu and Google Dengue Trends services [8] that employed server search logs.

Number of Diseases Monitored

Regarding the epidemiological focus of these works in terms of diseases of interest one can formulate two broad categories that is work that has concentrated on a singular illness such as flu-prediction.com, flusurvey.co.uk, Google Flu and Google Dengue Trends that concentrated on the flu and Dengue fever respectively in addition to most work on Twitter which has concentrated on the flu [16-19,23]. The second category involves systems that are able to handle multiple diseases. Depending on the underlying method this second category can be further subdivided into systems that handle a determinate set of diseases and generalized approaches that can potentially identify an unbounded number of ailments.

The former include systems such as the ontology and taxonomy based systems which as already discussed ultimately rely on text matching to extract predefined disease entities. A second more complex instance of this category involves topic modeling based approaches such as ATAM based on LDA [57]. As already noted these require some hyper parameters like alpha and beta parameters in this case which themselves require some prior assumptions of topic distributions the topics in this case being individual diseases. As already mentioned, further human mediation is required to determine if the resulting topics make sense. In this case the number of possible ailments is bounded by the number of topics or clusters set at training time whereupon new messages are classified using some distance measure like Euclidean distance. Furthermore, these approaches do not directly return ailments but rather clusters of messages with a high likelihood containing the same topics. The key limitation of employing such fixed definitions is that such systems are incapable of accommodating new health concerns such as emerging diseases without significant re-work.

For a more direct and truly generalized automated approach the most promising systems are those that apply machine learning based named entity recognition algorithms such as CRFs (55]. These rely on features like the order of words and other features like word prefixes and suffixes to label words as entities like persons, drugs and diseases. This approach is not limited to a reference dictionary and can potentially detect previously unseen ailments. A key drawback of this approach is that machine learning driven named entity recognition so far performs quite poorly on social data for instance the experiments by Jimeno-Yepes et al [55] reported a precision of 0.755 and a recall 0.62 on a data set of 11,647 tweets. This means that 25% of tokens identified as mentions diseases by the model were not and nearly 40% of tokens that referred to diseases were labeled as mentions of other entities. We would only expect performance to further deteriorate in a live implementation. However, we are not currently aware of any automated disease surveillance systems that have incorporated machine learning driven named entity recognition in their analytical pipeline.

As far as we know general threat monitoring has generally been deployed by human moderated systems like ProMed-Mail and GPHIN. Where the goal is predictive modeling, separate analytical pipelines are required for each new ailment as with Google Flu Trends for flu and Google Dengue Trends for Dengue fever as generally predictive epidemic models are not transferable between different diseases and from this perspective a generalized data extraction interface offers little gains however where the goal is to monitor some unspecified threat such as bioterrorism in real-time it would be extremely beneficial.

Predictive Modeling, Real-time Monitoring and Explanatory Modeling

Regarding the ultimate objective there are several ways to characterize the body of work, there are predictive systems that generally attempt to provide early warning for prospective disease outbreaks before they are reported by official systems, then there are systems intended primarily for monitoring the progress of outbreaks in real-time and finally there is work of an exploratory nature aimed at finding patterns and explaining disease related phenomena like causality, spatio-temporal clustering and disease dispersal networks that we refer to as explanatory modeling. Assuming a population of one individual figure 3 below depicts the usual progression of disease and the applicable system objectives at each phase in terms of prediction or monitoring from the point of view of that individual.

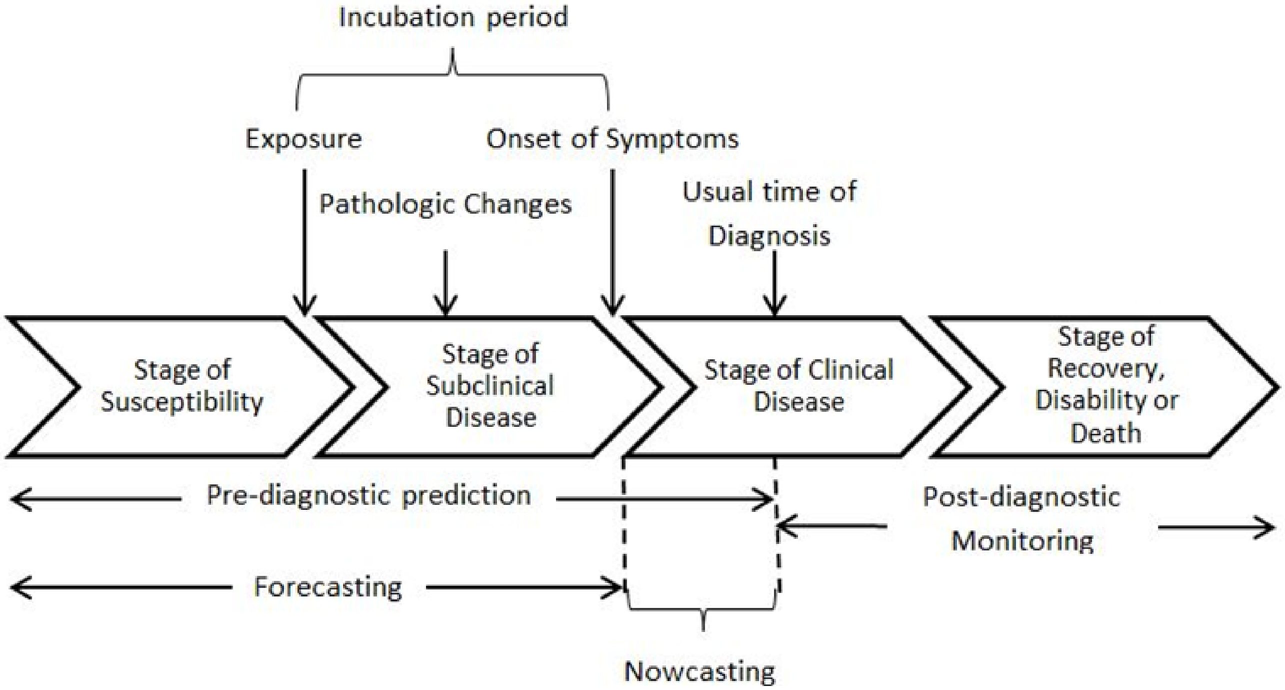
Incubation period
Exposure
Onset of Symptoms
Pathologic Changes
Usual time of Diagnosis
Stage of Susceptibility
Stage of Subclinical Disease
Stage of Clinical Disease
Stage of Recovery, Disability or Death
Pre-diagnostic prediction
Post-diagnostic Monitoring
Forecasting
Nowcasting

Figure 3: Disease forecasting, nowcasting and real-time monitoring vs. disease progression timeline

However, populations comprise several individuals and in general disease will progress differently for each individual with each phase commencing and concluding at different times. As a consequence there is significant overlap between prediction and monitoring. Where there is an unusually high number of individuals falling ill at roughly the same time for a given disease and population it is referred to as an outbreak.

In general we refer to implementations that employ pre-diagnostic data such as searches for certain keywords as predictive modeling and based on the prediction time horizon these can further be categorized into nowcasting and forecasting. Nowcasting refers to tracking outbreaks as they occur but due to process lags in formal surveillance systems like ILINet (Influenza-like Illness Surveillance Network) now-casting systems have been known to "predict" official statistics up to two weeks in advance [37]. Forecasting involves longer time horizons or at least is intended to for instance there are works that have employed the term forecasting for time horizons that are shorter or comparable to nowcasting systems for instance Dugas et al [65] use statistical methods to make predictions that are one week in advance of official statistics.

For a clearer technical distinction we can employ the shape of the model input data. Assuming a simple system with a single predictor variable x and a single target y and X and Y are vectors of values such that $X = (x_1, x_2, x_3 \dots \dots . x_n)$ and $Y = (y_1, y_2, y_3 \dots \dots . y_n)$. For nowcasting approaches such as the Google Flu Trends method the model input takes the form $\{(x_1, y_1), (x_2, y), (x_3, y_3) \dots \dots \dots . (x_n, y_n)\}$. Here the disease activity Y is assumed to be only dependent on the value of X at that given time and the goal is to predict Y given some X. For forecasting the model input is a time series corresponding to a chronological history of observed disease activity bounded by number of lag observations. For instance for a lag of two for data that is t observations long given Y_t denotes the value of Y at time t the model input would take the form $\{[(y_{t-n-2}), (y_{t-n-1}), (y_{t-n})] \dots \dots \dots \dots [(y_{t-4}), (y_{t-3}), (y_{t-2})], [(y_{t-3}), (y_{t-2}), (y_{t-1})]\}$ and the objective would be to predict Y at time t +1 given values of Y at time t and t - 1. In this case Y is dependent on the sequence of previous values of Y.

For the purpose of forecasting, a variety of supervised machine learning approaches have been applied and these include multivariate regression models [66], ensemble prediction approaches [67], Box-Jenkins Generalized Linear Models (GLM) and Generalized Linear Auto Regressive Moving Average (GARMA) [65]. Similar methods have been applied for nowcasting particularly regression techniques [64] in addition to aberrancy detection techniques [46]. For forecasting longest time horizons are around 2 months [68,69, 70]. However, the vast majority of work in this domain is on nowcasting and it entails systems like Google Flu and Google Dengue Trends, flunearyou.org and the majority of studies reviewed here [16-19].

Real-time disease monitoring is post-diagnostic meaning the condition has already been identified and it involves keeping track of suspected or confirmed outbreaks and epidemiological incidents. In these implementations there is no need to build predictive models to relate signal data to official data as these systems report cases with a high level of confidence and generally assume their sources to be true but not necessarily representative. These kinds of implementations will rely on high quality data sources like networks of public health professionals and include systems like HealthMap.org, ProMed-Mail, BioCaster, GPHIN and the "Outbreaks Near Me" mobile application. For implementations like the influenzanet.eu systems and flunearyou.org which rely on information provided directly by the population there is less certainty about the accuracy of diagnosis and therefore a weaker claim is made by employing more fluid syndromic incidence definitions rather than reporting cases of specific conditions. The general approach for such systems is to have users indicate their symptoms in an online form and if users exceed some number of symptoms the system takes this as a positive diagnosis for some syndrome as opposed to detecting a specific disease. For the systems mentioned above the syndrome is "flu-like illnesses".

Explanatory modeling is geared towards understanding the mechanics of disease emergence and spread and thereby identifying at risk individuals or populations and modeling effective interventions. With predictive modeling the mechanics are more or less irrelevant as all that is required is a positive correlative relationship between a signal variable that is usually the volume of messages about a given topic and disease activity on the ground. Explanatory modeling includes work on geospatial cluster analysis [47,68,70] and social networking analysis [71]. Geo-spatial cluster analysis involves determining clusters of locations with elevated level of cases during an outbreak. This is important because disease cases will not be uniformly geographically distributed and targeted interventions require more fine-tuned location data. Social networking analysis investigates interconnections between individuals and may be employed to determine at risk individuals or populations.

System Deployment Options

There are two main deployment options namely web and mobile. Most large scale deployments are in the form of web applications that run in a browser. The main advantage of this approach is that the service is made instantly available to most internet capable devices which normally ship with a default browser program. However, web implementations are not guaranteed to render uniformly as different vendors do not uniformly implement web standards like HTML (Hypertext Markup Language). In addition, the user experience for desktop sites when accessed from smaller devices is usually inferior to that on desktop due to the smaller size of the display. A common compromise is to create a minimal version of the web site referred to a mobile site. Web servers can then detect the client’s browser via information such as the “USER AGENT” HTTP (Hypertext Transfer Protocol) request header and load the appropriate website version. However, as depicted in figure 4 the number of users accessing mobile internet on smaller devices like smartphones and tablets has grown quickly since their introduction eventually overtaking that of users accessing internet via desktop on first November 2016. As a consequence of this there has been a rapid upsurge in mobile applications.

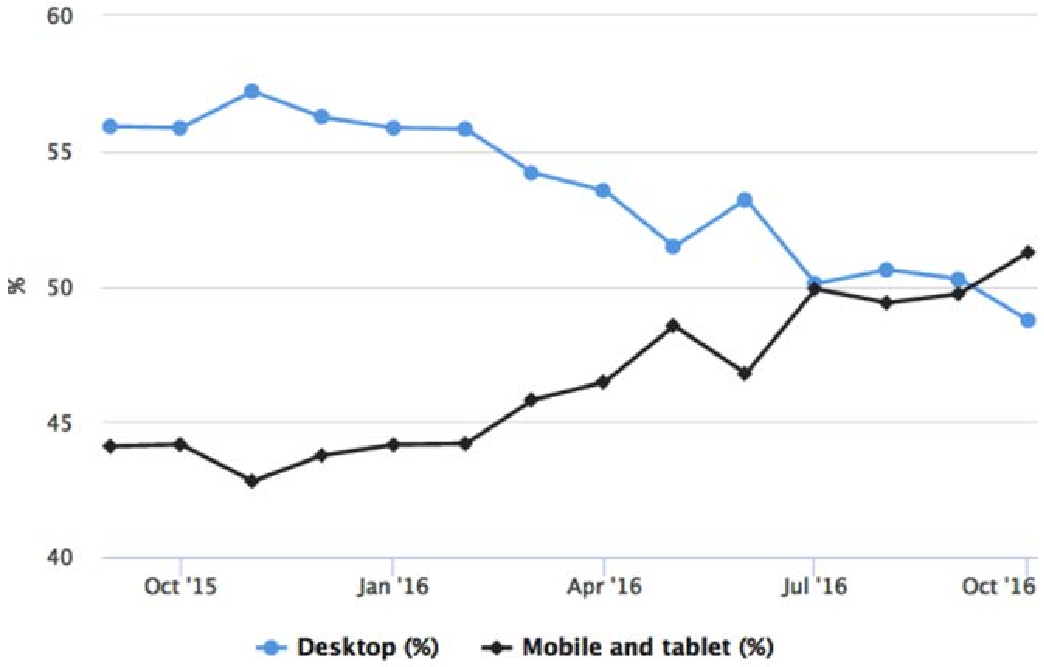

Figure 4: Proportion of worldwide internet users by access mode. Source: gs.statcounter.com [102]

Mobile applications have several advantages of web applications in terms of how they interact with the user especially where an explicit user participation model is employed. First of all most mobile applications do not explicit logon, in addition it is possible to program reminders and integrate them with the phone's native functions like phone vibration which could be more effective tools for maintaining user engagement than mass marketing campaigns such as email broadcasts that are typically employed by web applications. In addition, a more intimate connection is expected between participants and their phones which they keep on their person for most of the day than desktops or laptops which they interact with intermittently providing more opportunities for users to engage with the application for instance research shows that young adults use their smartphones an average of 5 hours a day which is 30% of their total waking time [72]. Interestingly, none of the large scale explicit user participation systems discussed so far offers a mobile website or mobile application for that matter. However, mobile website versions and applications are offered by Healthmap.org whose mobile platform "Outbreaks Near Me" is deployed on Android platform and Propeller Health that has both Android and IOS implementations. Both services employ an implicit user participation model.

Another important factor particularly for the developing world is price. Low end Android phones can go for as low as $25 USD [73] and can be used in areas with intermittent power or no power at all as they can be reliably recharged by off-grid options like solar chargers. They have also been applied extensively in monitoring diseases in plants and livestock where disease monitoring greatly benefits from mobile data collection units as it is impractical to transport plants and livestock to central facilities for diagnosis and treatment necessitating data collectors to actively seek information in the field. Examples include implementations like "Be seen Be safe" for real-time monitoring and modeling of poultry outbreaks in Canada which is available on both Android Playstore and Apple's app store and the Mcrops project by Makerere University in Uganda that employs image processing techniques to identify cassava diseases with smartphones and relay results in real-time to data analysts [74,75]. The Mcrops project has rolled out three apps on Android namely whitefly detection, Necrosis Detect and Adsurv. They are also particularly useful in fast changing environments such as the American navy's PASSION project (Precise At-Sea Ship System for Indoor-Outdoor Navigation system) which was intended to track on board outbreaks for the American naval fleet [76], the Android based system was an extension of previous work in Zambia and Colombia by the ONR (Office of Naval Research) and is applicable to dynamic situations like disaster relief in addition to the.

An important decision for mobile application deployments is the choice of platform as unlike web deployments different versions of the application are required for every target mobile platform. Developing for multiple platforms certainly imposes higher expenses and has to be clearly justified. Figure 5 depicts the popularity of different mobile platforms by market share from 2009 to 2017. As shown in the figure, the vast majority of devices are running Android operating system making it the default platform choice for most scenarios. In some countries like the United States the split between the two leading platforms namely Android and IOS is more even at 55.4% and 43% respectively [77] necessitating mobile implementations to roll out at least two versions of their applications to cater for the two options. For most large services the minimum deployment comprises a web application and at least one mobile implementation. The fact that several projects covered in this discussion have only a web deployment is indicative of the infancy of this domain.

IOS is generally very poorly represented in developing countries due to the fact that IOS products are typically more expensive. However, even the relatively low cost of Android devices may be prohibitive for many individuals in the poorest countries. In addition these devices require an internet subscription and internet coverage both of which cannot be taken for granted in many parts of the world. This has led to massive interest in SMS (Short Message Service) technology in the developing world. SMS based apps eliminate most of the financial hurdles associated with other technologies, in addition every single cell phone is SMS capable. The downside is that feature-wise SMS based interfaces are limited as they are text based. However, we are currently unaware of any efforts that have been implemented as truly social undertakings. The standard model seems to be to rely on a group of trained data entrants with a view of enhancing reporting within formal structures as opposed to sourcing health information directly from the general public [78-80].

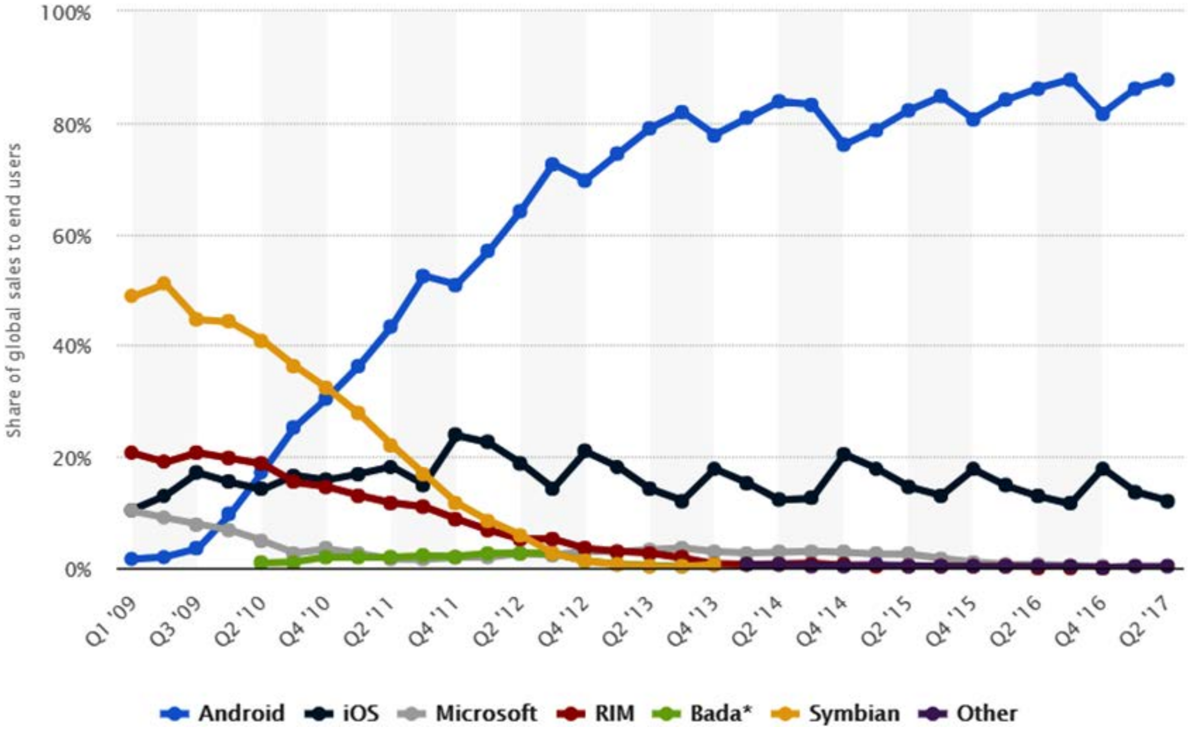

Figure 5: Mobile OS market share for different platforms. Source: Statista.com [103]

Summary

Whereas the preceding discussion has mainly focused on a textual analytical pipeline, we would like to point out that this is only because the majority of the work by far has employed textual signal data but it is not necessarily the limit of what is possible for instance Garimella et al [81] used automated image processing techniques like machine annotation to track lifestyle diseases such as obesity and excessive drinking using geo-tagged Instagram posts. We expect that as social media systems continue to grow in terms of users and machine learning and other analytical tools become more established more advanced forms of processing and data collection will become feasible. One such emerging field is the use of data from smart wearables and embedded sensors like the Propeller Health device referred to above that is attached to a user's inhaler or Bluetooth spirometer. In addition widely available devices like smart watches come packed with an array of sensors that given the right software can detect physiological signals like the wearer's heartbeat, muscle activity and stress levels. Embedded sensors make it possible to have high precision individual level surveillance of vital health signs like cardiac rhythms in a non-invasive way with very little conscious effort by the wearer. These technologies typically employ some intermediate computing device like a smart phone as a gateway node for local storage and data uplink to remote servers.

However, most vendors, with a few exceptions like Propeller Health, do not yet have dense enough networks of users to reliably execute public level health monitoring but certain product categories like smart watches are already widely deployed meaning we are quickly approaching a time when this will be feasible. Figure 6 shows the number of smart wearables sold worldwide by category from 2014 to 2017 depicting steady growth in the sector. It also reveals that majority of wearables are in the form of health and fitness trackers followed by smart watches. There is currently a lot of promising research in embedded sensor technologies capable of monitoring practically every physiological signal such as smart textiles that can measure stress levels and neural muscular abnormalities through embedded miniaturised ECGs (electrocardiograms) and EMGs (electromyograms) [82-84]. As the field advances these can only become cheaper, less invasive and more widespread.

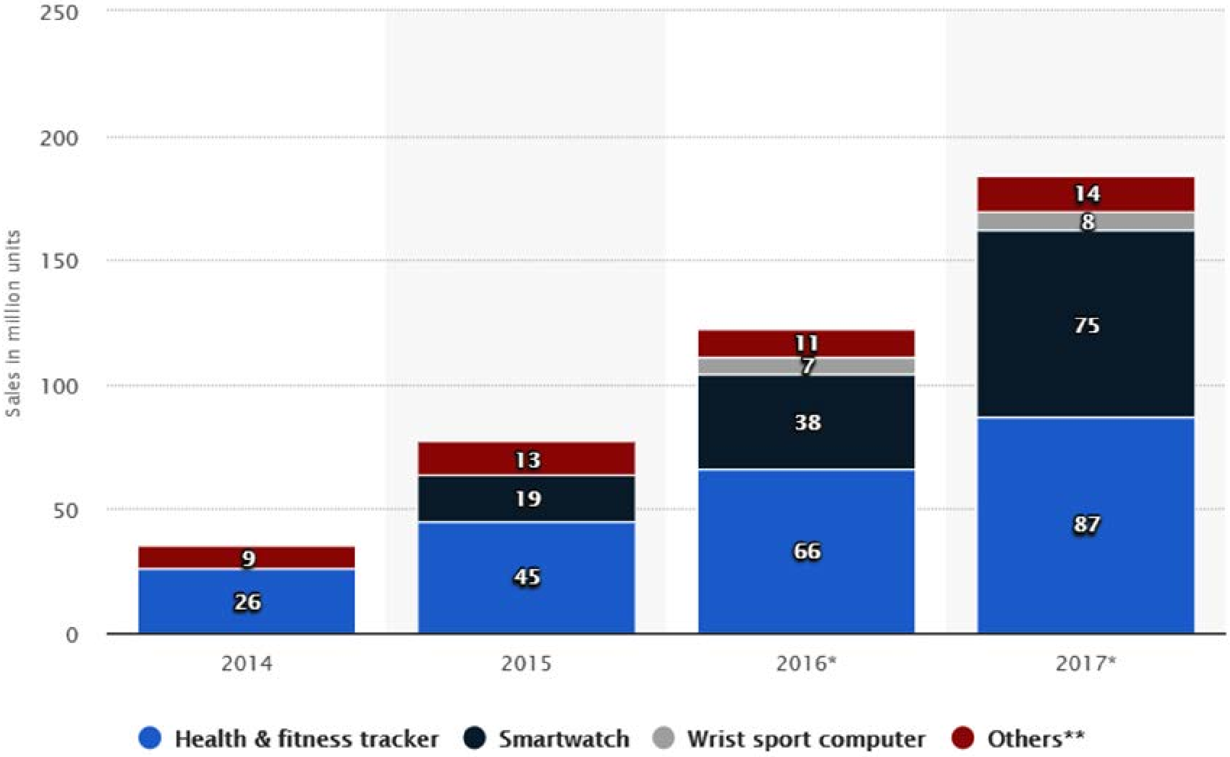

Figure 6: Yearly global sales of smart wearables from 2014 to 2017 by category from 2014 to2017. Source: Statista.com [104].

Finally we would like to point out that some systems mentioned in this discussion that are currently not live. These include Google Flu Trends and Google Dengue Trends that were retired in 2014, BioCaster which was retired in 2012 and Healthtweets.org and Crowdbreaks.org that are still in development. The Google Flu/Dengue Trends and BioCaster systems have been included because they were among the first large scale projects to be deployed for web based disease surveillance and their data and methods have widely been employed as benchmarks. Healthtweets.org and Crowdbreaks.org have been included because they are pioneering applications of large scale machine learning for textual processing in this domain. Table 1 below provides a summary of the preceding discussion.

Table I. A comparison of the merits and demerits of different design choices for automated disease surveillance on the social web

Design Element	***Design Option***	***Merits***	***Demerits***	***Examples in Literature (Live systems in bold, underlined, systems in development and prototypes in bold, retired systems in bold italics and studies in regular text)***
User Participation Model	Explicit User Participation	-Higher quality data -No text processing required	Limits potential pool of participants	**flunearyou.org** [23], **ProMed-Mail**, **influenzanet.eu** [24] (**flusurvey.org.uk**, **grippenet.fr**, **gripenet.pt**, **influenzasakkol.se**, **flusurvey.ie,influweb.it**)
	Implicit user Participation	All social media users are a potential source of data	-Requires advanced text processing -Potential for breaches of privacy	***Google Flu Trends and Google Dengue Trends*** [11], **Healthtweets.org** [20], **Crowdbreaks.org** [21], **GPHIN** [12], **flu-prediction.com,Healthmap.org** [16], **"Outbreaks Near Me"**, Several studies [16-19,43,44,81]
Language Parsing Methodology	Rule Based Parsing	Computationally inexpensive	Commonest approach of text matching is susceptible to false positives	***Google Flu Trends and Google Dengue Trends*** [11], ***BioCaster*** [45], **BioStorm** [46], **Healthmap.org** [16], **"Outbreaks Near Me"**, Several studies [11,19,41,43,44]
	Machine learning approaches	Better accuracy	-Computationally intensive -Model development is technically complex -Require labor intensive continuous generation and preparation of training data	**Crowdbreaks.org** [21], **Healthtweets.org** [20], Several studies [53-59]
Multiplicity of data sources	Single Data Source	low operational overhead in terms of data acquisition	Resulting models may be less robust than models built from multiple data sources	***Google Flu Trends and Google Dengue Trends*** [11], **Healthtweets.org** [20], **Crowdbreaks.org** [21],***BioCaster*** [45], **flu-prediction.com**, Several studies [16-19,23,44,62-64]
	Multiple Data Sources	-Models are likely to be more robust than those from a single data source	High operational overhead regarding data acquisition for instance maintaining multiple APIs	**Healthmap.org** [16], **"Outbreaks Near Me"**, **GPHIN** [12], **Pro-Med Mail**, Several studies [66,67]
System Deployment	Web deployment	-Single deployment may cater for multiple devices	-Application may not perform optimally across different devices	***Google Flu Trends and Google Dengue Trends*** [11], **Healthtweets.org** [20], **Crowdbreaks.org** [21], **BioCaster**

nt Options		since the application is accessible to any device with a browser	and Display and usability are poor on devices with smaller device -Shallower platform integration limiting access device hardware features and hence limiting user interaction	[45], **flu-prediction.com**, **Pro-Med Mail**, **Healthmap.org** [16]
	Mobile deployment	-Enjoy deeper platform integration on mobile devices allowing for a more intuitive user interface for instance on Android developers can take advantage of phone features like the speaker, accelerometer, vibration and GPS location	-Unlike web deployments each platform requires its own version of the application and therefore there can be considerable development effort if it is intended to serve multiple platforms	**Propeller health**(Android and IOS),**Outbreaks Near Me**(Android only)

Limitations of Automated Disease Surveillance Systems and Studies on the Social web

In this section we discuss some key caveats pertaining to the operationalization of disease surveillance based on social data. These are meant to inform implementers on the limits of these methods and guide deployment decisions and the interpretation of results. These include issues like the irregular distribution of signal and target data leading to geo-spatial gaps in data, inconsistencies in the location resolution of data, inherent data biases and the availability of data and language resources particularly for supervised machine learning method.

Geo-spatial Gaps in Data and Inconsistent Location Resolution of Data Points

Services like Google and Twitter are capable of obtaining a user's location with a resolution in excess of a millionth of latitude and longitude, an area equivalent to a thirty square feet wide circle on the ground, however most data is not geo-tagged and the distribution is such that most social web users are clustered into a few places like urban centres. Furthermore, although it is technically trivial to obtain high level location data like the country of origin of a social post with the leading social APIs, more fine grained geo-location usually requires users to explicitly disclose their location and most users do not. As a consequence most studies are based on data that is aggregated nationwide [11,16,17,19,44,66,67].

The problem with this is that it implicitly assumes that an entire nation or even a city is a single homogenous region. Such models are limiting as these predictions are ostensibly created to allow health officials to plan in advance but in this regard they fall short as they give no information on the geographical distribution of cases. Several types of heterogeneity do exist meaning such that it may be more useful to view the population as a metapopulation1. That is local conditions such as the immunity of different subpopulations and the efficacy of medical interventions may so significantly alter the parameters of a geographically distributed epidemic or outbreak that it may be more accurate to model it as a collection of distinct sub epidemics that may even be out of sync with each other. As an example figure 7 below shows district level epidemic trajectories for the 2014-2015 Ebola outbreak in Sierra Leone. It is quite clear that there were significant differences in epidemic curves for different districts in particular the curves for Kenema and Kailahun in the middle plot were significantly out of phase with other districts. Furthermore, even though the first cases were observed in Kenema and Bo the number of cases shot up rapidly in Kenema and rose at a much slower rate in Bo peaking as the incidence had already began to decline in Kenema. Such effects would be concealed by a country level model. Unfortunately for most locations it is currently not possible to reliably obtain sufficiently dense geo-tagged social data at this level of detail. However, where possible we recommend smaller models, such as district level models in this case, from which global models can be obtained by aggregation as demonstrated by Dugas et al [65] who used Google Flu Trends data to create predictions for individual medical centers. In most cases the distribution and availability of both signal and target data prohibits such an approach.

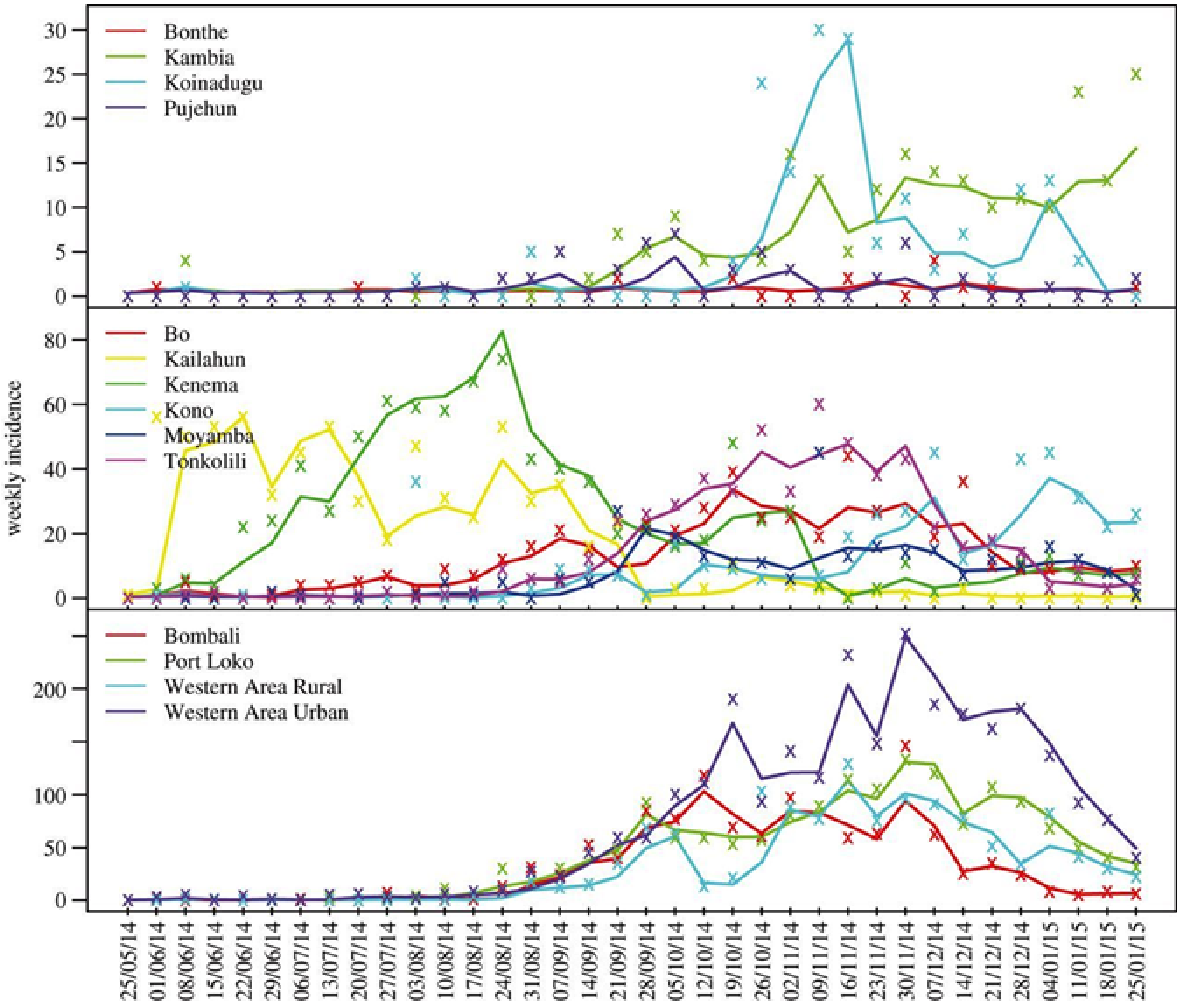

Figure 7: District level epidemic trajectories for the 2014-2015 Ebola outbreak in Sierra Leone. Weekly incidence records for each district are shown as colored 'x', solid line in the corresponding color is the approximate average incidence. Dates shown on the x-axis (dd/mm/yy) are endings of epidemic weeks. Source: Yang et al [105]

For the case of unavailable target data, work by Generous et al [68] has shown that in some situations models may be transferable between locations for particular diseases. They obtained positive results for influenza between Japan and Thailand and Thailand and the United States. This means that at least in the context of availability of target data it may be possible to employ models from resource-rich locations and contexts to at least make default, fallback predictions for resource-poor locations and contexts given demonstrably similar causal factors. However, in situations where it is not possible to obtain sufficiently dense social data for a given target location these methods are completely unusable.

Data Biases

The effect of some inherent biases has not been rigorously examined. Olteanu et al [85] highlight four possible biases namely population biases, behavioral biases, temporal variations and data redundancy. Population biases have to do with the distribution of different demographics, this is of importance here as demographics play some role in the spatio-temporal dynamics of disease spread [86]. This compounded with the fact that social data is often heavily tilted in favor of certain demographic categories has raised some queries about its representativeness. In some studies attempts have been made to correct these distortions for instance Tilston et al [39] split their data according to age category then re-weight it such that incidence frequencies for categories that were underrepresented in the data versus the population were upwardly adjusted whereas those for categories that were overrepresented in the data versus the population were downwardly adjusted. The weight of an individual $W_i$given P_i^{UK} is the actual proportion of the individual's population category in the population and $P_i^{Flusurvey}$ is the proportion of the individual's population in the survey data was then computed as follows:

$$W_i = P_i^{UK} / P_i^{Flusurvey} \quad (2)$$

However, systematic investigations into the role of these demographic distortions on the performance of these methods are generally lacking. For instance in the work by Tilston et al no empirical justification is made for weighting data such as comparing the performance of the weighted approach against that of an un-weighted approach.

Behavioral biases refer to how users behave in different contexts for instance Teevan et al [87] found that Twitter queries are more related to transient, temporary relevant situations whereas web queries by users about a topic evolved as users knowledge about topics increased. In other words, users had specialized purposes for different social services. Behavioral biases also include content production biases which encompass several observations such as the fact that there are geo-spatial linguistic variations within countries [88]. Temporal variations refer to time dependent differences in populations and behaviors. For instance if you carried out the same exact experiment at different times it is possible to sample entirely different users [89,90].

As a consequence of all this models eventually age and become inconsistent with the reality on the ground requiring new models to be created over time. Training new models over time will impose additional overheads which must be anticipated at design time. It is also important that experiments employ the correct temporal granularity or risk missing instances of relevant phenomena that is for short-lived phenomena a fine grained temporal granularity is required and conversely a coarse grained temporal granularity is required for long term phenomena. For instance you would have to be mindful of disease specific temporal characteristics like communicable periods and incubation periods. As an example the incubation period (the time between infection and first onset of symptoms) of Salmonella food poisoning is 6-48 hours therefore data for predictive modeling or monitoring of Salmonella outbreaks must originate from a data source that is updated at least hourly otherwise predictions and estimates may lag or become out of sync with the reality events on the ground.

Finally, redundancies may bias the data. It is not uncommon to have the same message posted several times or have several messages which are fundamentally equivalent. For instance Twitter has a “retweet” function which essentially allows users to propagate a message of interest throughout their network. Care has to be taken to deal with such redundancy or it may bias the quantitative inferences made from the data. Retweets are easy to deal with but where several messages refer to the same event but are lexically divergent and posted by different users advanced cross document co-reference resolution is required to determine entity mentions that are equivalent. We are presently unaware of any work on cross document co-reference resolution of disease event mentions on the social web. Matters become even more complicated where multiple data sources are employed as key events like disease outbreaks are likely to be communicated via several outlets. Redundancy does not only exist for major events but even at an individual user level as many users have multiple social accounts. This raises important questions about how best to architect multi-data source solutions for instance is it better to merge all the data sources into a single data stream or to create an ensemble of predictive models? Such questions are currently unanswered.

Accessibility of Data and Availability of Language Resources

Where supervised machine learning techniques are applied for language parsing there is need for annotated in-domain data. Given the sheer volume of messages on social media and the speed with which the lexicon evolves [11] creating representative datasets is very challenging. Generally speaking work on language processing disease surveillance in the social web has dealt with comparatively small datasets for instance Conway et al [91] use N-grams and semantic features to classify documents in the BioCaster corpus which comprises 1000 documents, Paul and Dredze [18] have to rely on Amazon’s Mechanical Turk (MTurk) tool [92] to generate a corpus of 5,128 tweets labeled as related or unrelated. These are relatively small compared to standard corpora like the IMDB corpus [93] which contains 50,000 annotated reviews. Small datasets increase the likelihood of over fitted models which are characterized by extremely high performance on training data but unreliable performance on production data.

Also there is a general lack of standard datasets and in most cases studies have had to resort to creating their own data. Other systems like Crowdbreaks.org have attempted to circumvent this problem by relying on crowd sourced message annotation by having site visitors label messages as either relevant or irrelevant. However, this comes at the price of having inconsistent definitions of message relevancy. Given the dynamic nature of social web data and users continuous and semi-automated means of corpora generation and maintenance are necessary. How to exactly arrive at a logistically optimal method for this however requires some dedicated research. In addition there is need for some dedicated effort towards creating standardized language resources which at the moment do not really exist making it difficult to compare the performance of different approaches to similar tasks. For text data the usual data source is server search logs [1,44,69], for social media data the majority of the work has employed Twitter's streaming API which allows free access to 1% of the Twitter feed. Another open option is Wikipedia article access logs [68]. However, in general terms free public access to data is uncommon and most systems require high level organizational approval for any form of meaningful access.

Overfitting

Where machine learning and statistical modelling are employed the practise is to sample as much data as possible and then split this data into training and testing datasets. The instinct is then to maximise the performance of whatever learned models on the training data however this makes two unjustifiable assumptions. The first is that the data is representative. This is problematic because of the huge amount of data involved for instance 400 million tweets only represents 1% the tweets in a year [94] and this percentage is ever getting smaller. However, as noted in the previous section, particularly with supervised approaches for language parsing the annotated training data is seldom more than a few thousand messages. It is therefore quite difficult to determine how much confidence may be assigned to such models. A second even more troublesome assumption is that the problem space is well defined. There are indeed unknown unknowns coupled with a dynamic problem space further complicating the modeling problem. Furthermore, researchers are generally more eager to report results on high performance models but in many instances these models essentially just overfit the training data. This means that models are tuned to maximize performance on the training sample rather than to generalize. A well known example of a model that performed well in development but poorly on live data is that of that of Google Flu Trends which was reported to have obtained a 0.97 mean correlation with actual data for the 2008-2009 flu American flu season in the initial paper [37] before grossly over-estimating the 2011-2012 and the 2012-2013 seasons [95,96].

Conclusion

Whereas the preceding section highlights several limitations, it is important to note that the caveats highlighted are not unique to the social web and are more or less general issues related to electronic disease surveillance but the sheer volume, variety and velocity of social data and the unprecedented reach of the social web amplify these problems to new proportions. However, those very characteristics of social data present previously unavailable opportunities for disease surveillance. For instance, for most social platforms such as social networking sites and microblogs this data is unsolicited essentially making the social web a self-updating database. This effectively eliminates the need for elaborate data collection infrastructure as data gathering is simply a matter of implementing the relevant social API and is essentially external to the surveillance process. Social APIs further simplify additional tasks like geo-location and information filtering which would otherwise extremely complex to operationalize with traditional approaches.

Furthermore many of these pitfalls like geo-spatial gaps in data and unrepresented demographics are effectively solving themselves as a consequence of contemporary technology and social trends for instance more and more people are accessing mobile computing devices and connecting to the internet [97,98]. Provided data exists in the form of signal data like search engine queries and social media messages and target data like ILINet statistics, web based techniques have been shown to have a wide range of applicability. Promising results have been obtained for most disease categories by mode of transmission. These include vector borne diseases [11], airborne diseases [16-19], food borne diseases

[63,64,99], contagious diseases [69], water borne diseases [62], sexually transmitted diseases [71], non communicable diseases [100] and even psychiatric conditions like suicide risk [101].

However, data remains a huge challenge. Whereas the social web alleviates many of the infrastructural burdens related with traditional formal systems it still requires target data for model development and validation. Unfortunately, in many contexts electronic health reporting is poorly developed making it impossible to build predictive models or validate the efficacy of social data for monitoring purposes. In addition, as noted earlier access to data like server search logs is usually heavily restricted therefore limiting the pool of researchers and systems that can employ it to those with high level organizational clearance. Regardless, major strides have been made with the little open data available and based on the enormous research interest and the evidence from the few large scale deployments we anticipate that many more stakeholders will begin to see this as an essential part of the disease surveillance process.

References

1. Buehler JW, Hopkins RS, Overhage JM, Sosin DM, Tong V. Framework for Evaluating Public Health Surveillance Systems for Early Detection of Outbreaks: Recommendations from the CDC Working Group, 2004. Available at https://www.cdc.gov/mmwr/preview/mmwrhtml/rr5305a1.htm, Accessed28-07-2017.

2. Gushulak BD, MacPherson DW. 2000. Population Mobility and Infectious Diseases: The Diminishing Impact of Classical Infectious Diseases and New Approaches for the 21st Century. *Clin Infect Dis*. 31(3), 776-80. doi:https://doi.org/10.1086/313998. PubMed

3. Greenberg AJ, Haney D, Blake KD, Moser RP, Hesse BW. Differences in Access to and Use of Electronic Personal Health Information Between Rural and Urban Residents in the United States. The Journal of Rural Health. 2016,p. n/a–n/a. Available from: http://dx.doi.org/10.1111/jrh.12228.

4. Seitio-Kgokgwe O, Mashalla Y, Seloilwe E, Chida N. Utilization of the District Health Information Software (DHIS) in Botswana: From paper to electronic based system. In: 2016 IST-Africa Week Conference, 2016. p. 1–10.

5. Jawhari B, Keenan L, Zakus D, Ludwick D, Isaac A, et al. 2016. Barriers and facilitators to Electronic Medical Record (EMR) use in an urban slum. *Int J Med Inform*. 94, 246-54. PubMed https://doi.org/10.1016/j.ijmedinf.2016.07.015

6. Halpin H, Tuffield MA. Standards-based, Open and Privacy-aware Social Web, 2010. Available at https://www.w3.org/2005/Incubator/socialweb/XGR-socialweb-20101206/, Accessed28-07-2017.

7. Chaffey D. Global social media research summary 2017,2017.Available athttp://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/, Accessed 28-07-2017.

8. Zephoria.com.TheTop20ValuableFacebookStatisticsUpdatedNovember2017,2017.Available at https://zephoria.com/top-15-valuable-facebook-statistics/,accessed 22-11-2017

9. European Commission. 2017.Availableathttps://ec.europa.eu/jrc/en/scientific-tool/medical-information-system, accessed 28-07-2017.

10. World Health organization. 2017.Availableathttp://www.who.int/ihr/alert_and_response/outbreak-network/en/, accessed 28-07-2017.

11. Google. 2017. Available at https://www.google.org/flutrends/about/, accessed 28-07-2017.

12. Mawudeku A, Blench M, Boily L, St John R, Andraghetti R, et al. The global public health intelligence network. Infectious Disease Surveillance, Second Edition. 2013, p.457–469.

13. Sumino K, Locke ER, Magzamen S, Gylys-Colwell I, Humblet O, et al. 2017. Use of a Remote Inhaler Monitoring Device to Measure Change in Inhaler Usewith Chronic Obstructive Pulmonary Disease Exacerbations. *J Aerosol Med Pulm Drug Deliv*. PubMed

14. Kim MS, Henderson KA, Van Sickle D. 2016. Using connected devices to monitor inhaler use in the real world. *Respiratory Drug Delivery*. 2016, 37-44.

15. Son J, Brennan PF, Zhou S. 2017. Correlated gamma-based hidden Markov model for the smart asthma management based on rescue inhaler usage. *Stat Med*. 36(10), 1619-37. PubMed

16. Culotta A. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. In: Proceedings of the First Workshop on Social Media Analytics. SOMA ’10. New York, NY, USA: ACM, 2010. p. 115–122.Availablefrom:http://doi.acm.org/10.1145/1964858.1964874.

17. Eiji A, Sachiko M, Mizuki M. Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. EMNLP ’11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011.p. 1568–1576. Available from: http://dl.acm.org/citation.cfm?id=2145432.2145600.

18. Paul MJ, Dredze M. 2012. A model for mining public health topics from Twitter. *Health*. 11, 16-6.

19. Shaman J, Karspeck A, Yang W, Tamerius J, Lipsitch M. Real-Time Influenza Forecasts during the 20122013 Season. In: Nature communications, 2013.

20. Dredze M, Cheng R, Paul MJ, Broniatowski D. HealthTweets.org: a platform for public health surveillance using Twitter. In: AAAI Workshop on the World Wide Web and Public Health Intelligence, 2014. p. 593–596.

21. Palmer S. Crowd breaks tracks disease trends through social media, 2017. Availableathttps://www.huck.psu.edu/content/crowdbreaks-tracks-disease-trends-through-social-media,accessed 28-07-2017.

22. http://www.flu-prediction.com, accessed 28-07-2017.

23. Chunara R, Aman S, Smolinski M, Brownstein JS. 2013. Flu Near You: An Online Self-reported Influenza Surveillance System in the USA. *Online J Public Health Inform.* 5(1). https://doi.org/10.5210/ojphi.v5i1.4456

24. Perrotta D, Tizzoni M, Paolotti D. Using Participatory Web-based Surveillance Data to Improve Seasonal Influenza Forecasting in Italy. In: Proceedings of the 26th International Conference on

World Wide Web. WWW ’17. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017. p. 303–310. Available from: https://doi.org/10.1145/3038912.3052670.[25]

25. https://www.degrotegriepmeting.nl/, accessed 28-07-2017.

26. Vandendijck Y, Faes C, Hens N. 2013. Eight years of the Great Influenza Survey to monitor influenza-like illness in Flanders. *PLoS One*. 8(5), e64156. PubMed https://doi.org/10.1371/journal.pone.0064156

27. Van Noort S, Muehlen M, Rebelod AH, Koppeschaar C, Lima LJ, and Gomes M. Gripenet: an internet-based system to monitor influenza-like illness uniformly across Europe. Euro surveillance: bulletin europeen sur les maladies transmissibles= European communicable diseasebulletin. 2007,12(7):E5–6.[28]

28. https://www.influweb.it/,accessed 28-07-2017

29. https://grippenet.fr/,accessed 28-07-2017.

30. http://www.flusurvey.org.uk, accessed 28-07-2017.

31. Brooks-Pollock E, Tilston N, Edmunds WJ, Eames KT.Using an online survey of healthcare-seeking behaviour to estimate the magnitude and severity of the 2009 H1N1vinfluenza epidemic in England. BMC infectious diseases.2011,11(1):68.[32], 2017

32. https://www.influensakoll.se/, accessed 28-07-2017.

33. https://flusurvey.ie/,accessed 28-07-2017

34. Kjelsø C, Galle M, Bang H, Ethelberg S, Krause TG. 2016. Influmeter–an online tool for self-reporting of influenza-like illness in Denmark. *Infect Dis*. 48(4), 322-27. PubMed https://doi.org/10.3109/23744235.2015.1122224

35. https://www.gripenet.es/,accessed 28-07-2017

36. http://www.promedmail.org,accessed 28-07-2017

37. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, et al. 2009. Detecting influenza epidemics using search engine query data. *Nature*. •••, 457. PubMed

38. Perrotta D, Tizzoni M, Paolotti D. Using Participatory Web-based Surveillance Data to Improve Seasonal Influenza Forecasting in Italy. In: Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017. p. 303–310.

39. Tilston NL, Eames KT, Paolotti D, Ealden T, Edmunds WJ. 2010. Internet-based surveillance of Influenza-like-illness in the UK during the 2009 H1N1 influenza pandemic. *BMC Public Health*. 10(1), 650. PubMed https://doi.org/10.1186/1471-2458-10-650

40. Land-Zandstra AM, van Beusekom MM, Koppeschaar CE, van den Broek JM, et al. Motivation and learning impact of Dutch flu-trackers. Journal of Science Communication.2016, 15.

41. Doan S, Ohno-Machado L, Collier N. Enhancing Twitter data analysis with simple semantic filtering: Example in tracking influenza-like illnesses. In: Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on. IEEE, 2012. p. 62–71.

42. Briscoe T, Carroll J, Watson R. The second release of the RASP system. In: Proceedings of the COLING/ACL on Interactive presentation sessions. Association for Computational Linguistics, 2006. p. 77–80.

43. de Quincey E, Kostkova P. Early Warning and Outbreak Detection Using Social Networking Websites: The Potential of Twitter. Kostkova P, editor. Berlin, Heidelberg: SpringerBerlin Heidelberg, 2010. Available from:https://doi.org/10.1007/978-3-642-11745-9_4.

44. Polgreen PM, Chen Y, Pennock DM, Nelson FD, Weinstein RA. Using Internet Searches for Influenza Surveillance. Clinical Infectious Diseases. 2008, 47(11):1443–1448. Available from:+http://dx.doi.org/10.1086/593098.

45. Collier N. Uncovering text mining: A survey of current work on web-based epidemic intelligence. In: Global public health, 2012.

46. Nyulas CI, O'Connor MJ, Tu SW, Buckeridge DL, Okhmatovskaia A, et al. An Ontology-Driven Framework for Deploying JADE Agent Systems.In:2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. vol. 2, 2008.p. 573–577.

47. Souza RC. Assunc ̧˜ao RM, de Oliveira DM, de Brito DE, MeiraJr W. Infection Hot Spot Mining from Social Media Trajectories. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2016. p. 739–755.

48. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. 2008. HealthMap:Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. *J Am Med Inform Assoc*. 15(2), 150-57. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2274789/#!po=79.4118. PubMed https://doi.org/10.1197/jamia.M2544

49. https://translate.google.com/intl/en/about/languages/,accessed 28-07-2017.

50. Collier N, Doan S, Kawazoe A, Goodwin RM, Conway M, et al. 2008. BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics*. 24(24), 2940-41. PubMed https://doi.org/10.1093/bioinformatics/btn534

51. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space .arXiv preprintarXiv:13013781. 2013

52. Magumba MA, Nabende P. An Ontology for Generalized Disease Incidence Detection on Twitter. In: International Conference on Hybrid Artificial Intelligence Systems. Springer, 2017. p. 38–51.

53. Lamb A, Paul MJ, Dredze M. Separating Fact from Fear: Tracking Flu Infections on Twitter. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia: Association for Computational Linguistics, 2013. p. 789–795. Available from: http://www.aclweb.org/anthology/N13-1097.

54. Doan S, Hung-Ngo Q, Kawazoe A, Collier N. Global Health Monitor - A Web-based System for Detecting and Mapping Infectious Diseases. In: IJCNLP, 2008

55. Jimeno-Yepes A, MacKinlay A, Han B, Chen Q. 2015. Identifying Diseases, Drugs, and Symptoms in Twitter. *Stud Health Technol Inform*. 216, 643-47. PubMed

56. Denecke K, Nejdl W. 2009. How valuable is medical social media data? Content analysis of the medical web. *Inf Sci*. 179(12), 1870-80. https://doi.org/10.1016/j.ins.2009.01.025

57. Michael JP, Mark D. 2014. Discovering health topics in social media using topic models. *PLoS One*. 9(8), e103408. PubMed https://doi.org/10.1371/journal.pone.0103408

58. Lu Y, Zhang P, Liu J, Li J, Deng S. 2013. Health-related hot topic detection in online communities using text clustering. *PLoS One*. 8(2), e56221. PubMed https://doi.org/10.1371/journal.pone.0056221

59. Chen AT. 2012. Exploring online support spaces: using cluster analysis to examine breast cancer, diabetes and fibromyalgia support groups. *Patient Educ Couns*. 87(2), 250-57. PubMed https://doi.org/10.1016/j.pec.2011.08.017

60. Keller M, Blench M, Tolentino H, Freifeld CC, Mandl KD, et al. 2009. Use of unstructured event-based reports for global infectious disease surveillance. *Emerg Infect Dis*. 15(5), 689. PubMed https://doi.org/10.3201/eid1505.081114

61. Frederic F. Google News: the secret sauce, 2013. Available at https://www.theguardian.com/technology/2013/feb/25/1, accessed 28-07-2017.

62. Chunara R, Andrews JR, Brownstein JS. 2012. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *Am J Trop Med Hyg*. 86(1), 39-45. PubMed https://doi.org/10.4269/ajtmh.2012.11-0597

63. YoussefAgha AH. Jayawardene WP, Lohrmann DK. Role of Social Media in Early Warning of Norovirus Outbreaks: A Longitudinal Twitter-Based Infoveillance. In: Proceedings of the International Conference on Data Mining (DMIN). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2013. p. 1.

64. Diaz-Aviles E, Stewart A. Tracking Twitter for Epidemic Intelligence. Case study: EHEC/HUS outbreak in Germany.2011, p. 82–85.

65. Dugas A, Jalalpour M, Gel Y, Levin S, Torcaso F, et al. 2013. Influenza Forecasting with Google Flu Trends. *Online J Public Health Inform*. 5(1). http://journals.uic.edu/ojs/index.php/ojphi/article/view/4470. PubMed https://doi.org/10.5210/ojphi.v5i1.4470

66. Shihao Y, Mauricio S, John SB, Josh G, Stewart R, SCK. Using electronic health records and Internet search information for accurate influenza forecasting. 2017 05,17.

67. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, et al. 2015. Combining search social media and traditional data sources to improve influenza surveillance. *PLOS Comput Biol*. 11(10), e1004513. PubMed https://doi.org/10.1371/journal.pcbi.1004513

68. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. 2014. Global disease monitoring and forecasting with Wikipedia. *PLOS Comput Biol*. 10(11), e1003892. PubMed https://doi.org/10.1371/journal.pcbi.1003892

69. Xu D, Liu Y, Zhang M, Ma S, Cui A, et al. Predicting epidemic tendency through search behavior analysis. In: IJCAI Proceedings-International Joint Conference on Artificial Intelligence. vol. 22, 2011. p. 2361.

70. Beswick A. # Outbreak: An Exploration of Twitter metadataas a means to supplement influenza surveillance in Canada during the 2013-2014 influenza season, 2016

71. Weibel N, Desai P, Saul L, Gupta A, Little S. HIV Risk on Twitter: The Ethical Dimension of Social Media Evidence-based Prevention for Vulnerable Populations. 2017

72. Andrews S, Ellis DA, Shaw H, Piwek L. 2015. Beyond self-report: tools to compare estimated and real-world smartphone use. *PLoS One*. 10(10), e0139004. PubMed https://doi.org/10.1371/journal.pone.0139004

73. BBC. Mozilla plans '$25 smartphone' for emerging markets, 2014.Available athttp://www.bbc.com/news/technology-26316265, accessed 27-12-2017.

74. Quinn J. Computational Techniques for Crop Disease Monitoring in the Developing World. In: International Symposium on Intelligent Data Analysis. Springer, 2013. P.13–18.

75. Mwebaze E, Schneider P, Schleif FM, Aduwo JR, Quinn JA, et al. 2011. Divergence-based classification in learning vector quantization. *Neurocomputing*. 74(9), 1429-35. https://doi.org/10.1016/j.neucom.2010.10.016

76. DOD. ONR Program Uses Cel lPhones to Fight Epidemics, 2013.Availableathttp://science.dodlive.mil/2013/02/03/onr-program-uses-cell-phones-to-fight-epidemics/,accessed 27-12-2017.

77. Gs.statcounter.com. Mobile Operating System Market Share in United States of America - November2017,2017.Availableathttp://gs.statcounter.com/os-market-share/mobile/united-states-of-america, accessed 27-12-2017.

78. Toda M, Njeru I, Zurovac D, Tipo SO, Kareko D, et al. 2016. Effectiveness of a Mobile Short-Message-Service–Based Disease Outbreak Alert System in Kenya. *Emerg Infect Dis*. 22(4), 711. PubMed https://doi.org/10.3201/eid2204.151459

79. Mwabukusi M, Karimuribo ED, Rweyemamu MM, Beda E. 2014. Mobile technologies for disease surveillance in humans and animals. *Onderstepoort J Vet Res*. 81(2), 1-5. PubMed https://doi.org/10.4102/ojvr.v81i2.737

80. National Center for Disease Control DGoHS. 2017. Available at http://idsp.nic.in/index4.php?lang=1&level=0&linkid=408&lid=3691&Background=Light&font=Decrease, accessed 27-12-2017.

81. Garimella VRK, Alfayad A, Weber I. Social media image analysis for public health. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, 2016. p. 5543–5547.

82. Schuchert A, Behrens G, Meinertz T. 1999. Impact of Long-Term ECG Recording on the Detection of Paroxysmal Atrial Fibrillation in Patients After an Acute Ischemic Stroke. *Pacing Clin Electrophysiol*. 22(7), 1082-84. PubMed https://doi.org/10.1111/j.1540-8159.1999.tb00574.x

83. Lobodzinski SS, Laks MM. 2012. New devices for very long-term ECG monitoring. *Cardiol J*. 19(2), 210-14. PubMed https://doi.org/10.5603/CJ.2012.0039

84. Leone A, Rescio G, Caroppo A, Sicilian P. 2015. A Wearable EMG-based System Pre-fall Detector. *Procedia Eng*. 120, 455-58. https://doi.org/10.1016/j.proeng.2015.08.667

85. Olteanu A, Carlos C, Fernando D, Emre K. Social Data:Biases, Methodological Pitfalls, and Ethical Boundaries,2016.Available at: https://ssrn.com/abstract=2886526, accessed 28-07-2017

86. Sloan C, Moore M. 2011. t Hartert. Impact of pollution, climate, and socio demographic factors on spatiotemporal dynamics of seasonal respiratory viruses. *Clin Transl Sci*. 4(1). PubMed https://doi.org/10.1111/j.1752-8062.2010.00257.x

87. Teevan J, Ramage D, Morris MR. TwitterSearch: A comparison of microblog search and web search. In: Proceedings of the Forth International Conference on WebSearch and Web Data Mining, WSDM 2011, 2011.

88. Mocanu D, Baronchelli A, Perra N, Goncalves B, Zhang Q, et al. 2013. The Twitter of babel: Mapping world languages through microblogging platforms. *PLoS One*. 8(4). PubMed https://doi.org/10.1371/journal.pone.0061981

89. Lampe C, Ellison NB, Steinfield C. Changes in use and perception of Facebook. In: In Proceedings of the ACM 2008 conference on Computer supported cooperative. ACM, 2008.p. 721–730.

90. Liu Y, Kliman-Silver C, Mislove A. The Tweets They Area-Changin: Evolution of Twitter Users and Behavior. In: International AAAI Conference on Web and Social Media Eighth International AAAI Conference on Weblogs and Social Media. AAAI, 2014.

91. Conway M, Doan S, Kawazoe A, Collier N. 2009. Classifying disease outbreak reports using n-grams and semantic features. *Int J Med Inform*. 78(12), 47-58. PubMed https://doi.org/10.1016/j.ijmedinf.2009.03.010

92. Callison-Burch C, Dredze M. Creating Speech and Language Data with Amazon’s Mechanical Turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with

Amazon’s Mechanical Turk. CSLDAMT ’10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. p. 1–12. Available from: http://dl.acm.org/citation.cfm?id=1866696.1866697.

93. Maas AL, Daly RE, Pham PT, Huang D, Ng AY, et al. Learning Word Vectors for Sentiment Analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. HLT ’11. Stroudsburg, PA,USA: Association for Computational Linguistics, 2011.p. 142–150. Available from: http://dl.acm.org/citation.cfm?id=2002472.2002491.

94. Godin F, Vandersmissen B, De Neve W, Van de Walle R. Named entity recognition for Twitter microposts using distributed word representations, 2015. ACL W-NUT NER shared task. ACL Workshop on Noisy User-generated Text. 2015

95. Lazer D, Kennedy R, King G, Vespignani A. 2014. The Parable of Google Flu: Traps in Big Data Analysis. *Science*. 343(6176), 1203-05. doi:https://doi.org/10.1126/science.1248506. PubMed

96. Butler, D. When Google got flu wrong. 2013. Nature. 494: 155–156. doi:https://doi.org/10.1038/494155a. PubMed

97. Poushter J. Smartphone ownership and internet usage continues to climb in emerging economies. Pew Research Center. 2016,22

98. deChoudhury M, Morris MR, White RW. Seeking and Sharing Health Information Online: Comparing Search Engines and Social Media. In: Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems. CHI ’14. New York, NY, USA: ACM, 2014. P.1365–1376 Available from http://doi.acm.org/10.1145/2556288.2557214.

99. Zhou Xc. 2010. ShenHb. Notifiable infectious disease surveillance with data collected by search engine. *Journal of Zhejiang University-Science C*. 11(4), 241-48. https://doi.org/10.1631/jzus.C0910371

100. Dai H, Lee BR, Hao J. 2017. Predicting Asthma Prevalence by Linking Social Media Data and Traditional Surveys. *Ann Am Acad Pol Soc Sci*. 669(1), 75-92. https://doi.org/10.1177/0002716216678399

101. Hagihara A, Miyazaki S, Abe T. 2012. Internet suicide searches and the incidence of suicide in young people in Japan. *Eur Arch Psychiatry Clin Neurosci*. 262(1), 39-46. PubMed https://doi.org/10.1007/s00406-011-0212-8

102. Gs.statcounter.com. Internet usage exceeds desktop for first time worldwide, 2017.Availableathttp://gs.statcounter.com/press/mobile-and-tablet-internet-usage-exceeds-desktop-for-first-time-worldwide,accessed 27-12-2017.

103. Statista.com. Global mobile OS market share in sales to end users from 1st quarter 2009 to 2nd quarter 2017, 2017. Available at https://www.statista.com/statistics/266136/global-market-share-held-by-smartphone-operating-systems/, accessed 27-12-2017.

104. Statista,com. Global mobile OS market share in sales to end users from 1st quarter 2009 to 2nd quarter 2017, 2017. Available athttps://www.statista.com/statistics/757623/wearables-sales-by-category-worldwide/,accessed 27-12-2017.

105. Yang W, Zhang W, Kargbo D, Yang R, Chen Y, et al. 2015. Transmission network of the 2014–2015 Ebola epidemic in Sierra Leone. *J R Soc Interface*. 12(112), 20150536. PubMed https://doi.org/10.1098/rsif.2015.0536

106. Gomide J, Veloso A, Meira W, Jr, Almeida V, Benevenuto F, et al. Dengue Surveillance Based on a Computational Model of Spatio-temporal Locality of Twitter. In: Proceedings of the 3rd International Web Science Conference. WebSci '11. New York, NY, USA: ACM, 2011.p. 3:1–3:8. Available from:http://doi.acm.org/10.1145/2527031.2527049.

107. Tseng KC, Lin BS, Liao LD, Wang YT, Wang YL. 2014. Development of a wearable mobile electrocardiogram monitoring system by using novel dry foam electrodes. *IEEE Syst J*. 8(3), 900-06. https://doi.org/10.1109/JSYST.2013.2260620

[1] A population of populations, it comprises members of the same species and a degree of separation is assumed between the different sub populations. Some level of interaction occurs as a result of some individuals moving between different subpopulations.